

Mobile pupillometry in manual assembly: A pilot study exploring the wearability and external validity of a renowned mental workload lab measure.

Bram B. Van Acker^{abc*}, Klaas Bombeke^c, Wouter Durnez^c, Davy D. Parmentier^a, João Costa Mateus^a, Alessandro Biondi^a, Jelle Saldien^{ac}, Peter Vlerick^b

^a Department of Industrial Systems and Product Design, Faculty of Engineering and Architecture, Ghent University, Technologiepark Zwijnaarde 46, 9052, Zwijnaarde, Belgium

^b Department of Personnel Management, Work and Organizational Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Henri Dunantlaan 2, 9000, Ghent, Belgium

^c Research group IMEC-MICT-Ghent University, De Krook, Miriam Makebaplein 1, 9000 Ghent, Belgium

* Corresponding author: bramb.vanacker@ugent.be, +32 478 57 14 11, orcid.org/0000-0002-6565-3569

Abstract: Human operators in the upcoming Industry 4.0 workplace will face accelerating job demands such as elevated cognitive complexity. Unobtrusive objective measures of mental workload (MWL) are therefore in high demand as indicated by both theory and practice. This pilot study explored the wearability and external validity of pupillometry, a MWL measure robustly validated in laboratory settings and now deployable in work settings demanding operator mobility. In an ecologically valid work environment, 21 participants performed two manual assemblies - one of low and one of high complexity - while wearing eye-tracking glasses for pupil size measurement. Results revealed that the device was perceived as fairly wearable in terms of physical and mental comfort. In terms of validity, no significant differences in mean pupil size were found between the assemblies even though subjective mental workload differed significantly. Exploratory analyses on the pupil size when attending to the assembly instructions only, were inconclusive. The present work suggests that current lab-based procedures might not be adequate yet for in-the-field mobile pupillometry. From a broader perspective, these findings also invite a more nuanced view on the current validity of lab-validated physiological MWL-measures when applied in real-life settings. We therefore conclude with some key insights for future development of mobile pupillometry.

Keywords: mental workload, pupillometry, assembly, external validity, wearability, eye tracking glasses

Author final peer reviewed version

1. Introduction

As the industrial workplace is steadily changing from a routine work environment to a small-batch, knowledge driven industry 4.0 workplace (cf., Shalin et al. 1996; Young et al. 2014; Brolin et al. 2017; Longo et al. 2017), the demand for operator mental workload (MWL) optimization grows prominently. To organize work as a function of optimal mental workload, work designers want to gain insight into the cognitive processes and the associated physiological reactions induced by specific mental workload antecedents such as task complexity, task switching, instruction format, human-cobot communication, etc. (Hoedt, Claeys, Van Landeghem, & Cottyn, 2017; Parmentier, Van, Detand, & Saldien, 2019; Stork & Schubö, 2010; Thorvald, Lindblom, & Andreasson, 2017; Van Acker, Parmentier, Vlerick, & Saldien, 2018; Young et al., 2014). If not, suboptimal mental workload can lead to errors, safety risks, and detrimental effects on operators' mental and physical wellbeing (Matthews 2016; Brolin et al. 2017; Kolbeinsson et al. 2017; Wickens 2017).

In order to measure mental workload, the human factors and ergonomics field builds on a strong tradition of subjective measures and (secondary task) performance monitoring (for an overview, see Young et al. 2014). These methods indirectly infer MWL through, respectively, operator perceptions and, for example, quality rate. Other accounts provide researchers and practitioners with more direct methods gauging the physiological resources being spent when attending to tasks (for an overview, see Charles and Nixon 2019). Theta brainwaves and suppression of alpha brainwaves measured with wearable electroencephalography (EEG) headsets, for instance, have shown to be indicators of operator MWL (Wilson and Russell 2003; Ryu and Myung 2005; Antonenko et al. 2010; Huang et al. 2013; Hairston et al. 2014; Guru et al. 2015), while functional Near-infrared Spectroscopy (fNIRS) measures cerebral blood flow velocity at the frontal brain regions to estimate MWL (Ayaz et al., 2013, 2012; Foy & Chapman, 2018; Howard et al., 2015; McKendrick et al., 2016). These measures might however be most promising in fixed occupational positions (e.g., air traffic control centers, nuclear plant control rooms, motor vehicles, trains, airplanes, etc., Young et al. 2014; for an example, see Wanyan et al. 2018) as they could be perceived as obtrusive and unsafe when worn in more mobile industrial contexts such as assembly work places. Other less obtrusive gauges, such as heart-rate parameters and electrodermal response (EDR) (Boucsein, 2012; Collet, Salvia, & Petit-Boulanger, 2014; Kocielnik, Sidorova, Maggi, Ouwerkerk, & Westerink, 2013; Serino, Matic, Giakoumis, Lopez, & Cipresso, 2016) could provide one solution for such environments (e.g., when integrated into a wristband) but require more research and development on its applicability. Research on heart-rate parameters, for instance, showed mixed results (cf., Heine et al. 2017), sometimes with parameters being invalid and not sensitive to subtle fluctuations in mental workload (see, e.g.,

Paas and Van Merriënboer 1994), while EDR has, to our knowledge, not yet been robustly validated in mobile settings. A final strongly validated MWL-measure that *is* sensitive to fluctuating MWL levels (Paas, Tuovinen, Tabbers, & van Gerven, 2003) is pupillometry, and interestingly, a measure that can now be deployed in mobile situations such as assembly work places.

1.1. Pupillometry

Since Hess and Polt (1960), pupil size has not only been related to changes in light conditions, but also to changes in cognitive and emotional load. When cognitive task demands increase, the pupil becomes larger, this, in a wide range of cognitive tasks such as solving mathematical equations (Kahneman, Tursky, Shapiro, & Crider, 1969), remembering a number of digits (Beatty & Kahneman, 1966; Peavler, 1974), Stroop interference (Laeng, Ørbo, Holmlund, & Miozzo, 2011), retrieving information from memory (van Rijn, Dalenberg, Borst, & Sprenger, 2012), attentional allocation (Verney, Granholm, & Marshall, 2004), visual search (Porter, Troscianko, & Gilchrist, 2007), cognitive control (van der Wel & van Steenbergen, 2018), etc. Peavler (1974) interestingly also found that pupil dilation can reach an asymptote when the limits of working memory are exceeded, while Granholm and colleagues (1996) discovered that when demands are too high, pupil diameter can in fact decrease again, thus showing a curvilinear relation with task demands. The last two decades showed a remarkable revival in pupillometry research because of the cost-efficiency and precision of pupillometry technology (see, e.g., Sirois and Brisson 2014; Mathôt 2018; van der Wel and van Steenbergen 2018). Also in the ergonomics and human factors field pupillometry has shown its potential in various contexts. Recarte and Nunes (2003) and Tsai et al. (2007) observed that the pupil diameter increases when engaging participants in more demanding secondary tasks while driving an automobile. Air traffic controllers revealed to have larger pupils under increased task difficulty (cf., Ahlstrom and Friedman-berg 2006; Truschzinski et al. 2018) and pupil sizes in participants in a simulated piloting task were larger when task demands increased (Causse, Peysakhovich, & Fabre, 2016). Based on real-time pupil dilation data, Katidioti et al. (2016) even developed a task-independent interruption management system (IMS) while simulating the job context of an employee in an electronics company answering to emails and being interrupted by chat messages. In a gaze interactive assembly instruction, Hansen et al. (2018) observed a pupil size decrease throughout a LEGO building process, except when the child participants had to back-step. Finally, differences between experts and non-experts in surgical skill have been related to pupil size (Erridge, Ashraf, Purkayastha, Darzi, & Sodergren, 2017; Richstone et al., 2010) and also harder visual-motor aiming tasks (such as with a one-handed

grasper in endoscopic surgery) yield higher pupil dilation (Jiang, Zheng, Bednarik, & Atkins, 2015; see also, Dalveren, Cagiltay, Ozcelik, & Maras, 2018).

Pupillometry is often restricted to stationary laboratory settings or applied (simulated) pilot, driving or air traffic controller settings requiring subjects to remain in a fixed position and engage in a limited field of vision - since remote eye-tracking systems otherwise lose track of the eyes (cf., Recarte and Nunes, 2003). Marinescu et al. (2018) even suggest their mixed-method approach including pupillometry, given the current technological possibilities, to be best suited for workplaces where employees are mainly seated. However, head-mounted eye-tracking systems, as for example used in driving research (cf., Tsai et al. 2007), allows applied pupillometry research to take one step further into the direction of fully mobile mental workload measurement. A recent technological development now even made this eye-tracking method more light-weighted and small-scaled by integrating eye-tracking cameras into the frame of a pair of glasses, providing researchers and practitioners a way to measure pupil size in real mobile work settings where users are fairly free to move, gaze into all directions and change positions around the work place (see similar applications for eye-tracking purposes in, e.g., Vansteenkiste et al. 2017). In the current study, participants wore such eye-tracking glasses (ETG) in a simulated assembly setting in which they, although being seated, changed positions between presentation of instructions and work table, inspected the assembly components in a naturalistic field of vision, then manually interacted with the components, and placed and screwed them together.

The main goal was to explore the wearability (in terms of physical and mental comfort, see below) and the external validity of pupillometry in a more mobile, naturalistic assembly environment in which, per definition, there is less control over confounding variance caused by the wide variety of cognitive, emotional and physical processes involved, such as operator motor control, non-restricted eye movement patterns at various eye-to-stimulus distances, but also factors external to the operator such as changes in luminance coming from different types of stimuli (instead of equiluminance standards in lab experiments, see, e.g., Braem et al. 2015; Bombeke et al. 2016) and other unknown error variance. Indeed, all these factors are inherent to measurement in real-life, but as well inherent to the limitations of pupillometry and other wearable MWL-measures. If the next step from lab measurement to field measurement is to be taken, thorough understanding of these factors and their inclusion into measurement protocols are imperative for sound validation. Our study was therefore conceptually underpinned as both task complexity and pupillometry were embedded within the latest MWL-framework of Van

Acker et al. (2018), implying that the study captured and controlled for some key confounding MWL related variables such as emotional load, physical load and visual-spatial intelligence.

Contrary to classic pupillometry studies inducing very specific types of MWL separately - e.g., arithmetic processing or visual scanning - the current study design moreover was ecologically valid as it allowed for a wider, more real-life spectrum of cognitive processing. We did so by designing a representative industrial translation of traditional laboratory MWL-manipulations in the form of two assemblies - one of low complexity and one of high complexity. Participants performed both tasks and had to manually select, orient and screw all parts together, based on a set of prescribed instructions. We then assessed wearability of the ETG and subjective MWL and took pupil size measures for both the entire assemblies as well as for each of the separate consecutive steps within both assembly tasks.

Altogether, the present pilot study is, to our knowledge, the first to apply pupillometry in an ecologically valid assembly setting - a setting ever more relevant as the assembly industry is convincingly transforming into an Industry 4.0 (Longo et al., 2017) work context in which job complexity strongly increases due to the required operator adaptivity and top-notch technological support. Indeed, as the smart factory of the future is being equipped with cyber-physical systems, Internet of Things, and Services, employers want and need to profoundly understand the mental load experienced by the crucial human operators in this environment. Objective physiological measures are pivotal to these ends. As pupillometry has shown to be a strong indicator of MWL under controlled, equiluminant laboratory settings restricting to a limited field of view, the present pilot study endeavors to explore for the first time how feasible the use of this physiological marker could be when measured with recently developed eye-tracking glasses and in a low-controllable naturalistic industrial environment.

1.2. Study overview

In a within-subjects experimental design, participants engaged in an assembly task of high and low complexity. For both assemblies, they were first shown pictures of the required end goal, were then presented with the instructions for one step and were asked to execute this step. After finishing the step, they continued with the following step, until the seventh and final step of the assembly was completed.

Following the triangulation method as proposed by Van Acker et al. (2018) (see also Matthews 2016), we also pinpointed the subjective perceptions following completion of both assemblies

next to pupil size. This is important, because mental workload is the result of a wide range of interacting variables and subjective measures can help in interpreting physiological data while, interestingly, they do not necessarily converge with them (cf., Brookhuis and De Waard 1993; Myrtek et al. 1994; Annett 2002; Young et al. 2014; Matthews et al. 2015). For these reasons, we asked participants after each assembly task to rate the mental work demands they experienced. We hypothesized that:

- H1: Experienced mental work load will be higher in the high complexity condition, as compared to the low complexity condition.

Based on the laboratory and applied pupillometry experiments as discussed above and having designed the high complexity assembly not to cause mental overload (cf., the mentioned curvilinear relation), we hypothesized that:

- H2: The high complexity condition, as compared to the low complexity condition, will induce a larger pupil size when averaging pupil diameter over the entire assembly task.

In line with the previous hypothesis and aiming to explore the sensitivity of the gauge (Matthews, Reinerman-Jones, Wohleber, et al., 2015) we compared the separate assembly steps and formulated the following hypothesis :

- H3: In each of the consecutive assembly steps, a larger pupil size will be triggered in the high complexity condition, as compared to the low complexity condition.

2. Methods

2.1 Participants

Some classic laboratory experiments leverage sample sizes in the range of, e.g., $N = 12$ (Porter et al., 2007), $N = 15$ (Piquado, Isaacowitz, & Wingfield, 2010) and $N = 19$ (van Rijn et al., 2012), while more applied fixed position experiments run trials with sample sizes going from $N = 6$ (Ahlstrom & Friedman-berg, 2006; Recarte & Nunes, 2003) to $N = 24$, $N = 25$ (Causse et al., 2016; Truschzinski et al., 2018). Because this study was set out as a pilot study exploring the feasibility of mobile pupillometry, we therefore decided to collect data of at least twenty participants as to stay in line with previous laboratory studies and to stay within the restrictions of what is feasible in the field.

A total of 21 university student volunteers and staff affiliated with an engineering university faculty (33.3% female, $M_{\text{age}} = 23.3$, $SD_{\text{age}} = 3.25$), naïve to the manipulation (afterwards, none of the participants reported to have an idea about the experimental goal), participated after giving their written consent. Since the eye-tracking glasses used cannot be worn on top of corrective glasses, people who could not read a computer screen from normal distance and would not be able to assemble in the present setting without glasses could not participate. People with corrective lenses or corrected to normal vision could still take part in the study (no participant reported to have had problems with their vision affecting execution of the assembly on a 7-point Likert scale, $M = 2.67$, $SD = 1.46$, $Min = 1$, $Max = 4$). Because of a technical failure with the pupillometry equipment we ended up with data of 19 participants for the pupil size analyses.

2.2 Experimental design

The study followed a within-subjects design: all participants performed both assemblies. The order of the assemblies was randomly counterbalanced. Participants were randomly assigned to one of the two orders and had a resting phase in between. In total, the experiment would take approximately one hour.

2.3 Designing assembly complexity

In order to simulate a real-life industrial translation of the standard laboratory MWL-manipulations such as n-back tasks and digit span tasks (cf., García et al. 2017; or for an overview, see van der Wel and van Steenbergen 2018), we designed an assembly of high and one of low complexity. We did so based on the task variables predicting assembly difficulty as proposed by Richardson et al. (2004, 2006), extended with Pillay's (1997) experimental approach of using non-familiar assembly objects (see also Norman's, 1983, Shalin et al.'s, 1996, insights on mental models in objects). We also added one material-related task variable manipulating complexity. Table 1 presents a total of 10 manipulated task variables and how they fed into the design of both assemblies. In so doing, we aimed to manipulate as much of the differential cognitive processes involved in typical assembly tasks as possible, such as perception (e.g., of part features), mental rotation (e.g., required orientation of parts), response selection (decision making) and action (e.g., motor execution) in commissioning and joining (cf., Shalin et al. 1996; Pillay 1997; Stork and Schubö 2010; Van Acker et al. 2018). We then tested and redesigned the composition of both assemblies, the display of the parts on the work table and the assembly instructions in an iterative design process with subject matter experts (SMEs) along a think aloud approach (cf., Altman 2007). The final iteration and pilot test confirmed that both assemblies were sufficiently distinctive when the SMEs rated the complexity of both assemblies on 7-point Likert-scales when

asked for a comparison score, how easy to hard and recognizable the assembly would be for the average operator and how hard the selection of the parts, mental rotation and orientation on the assembly-in-progress, and finding the correct fastening points would be for the average operator.

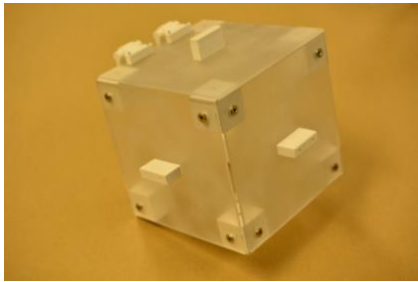
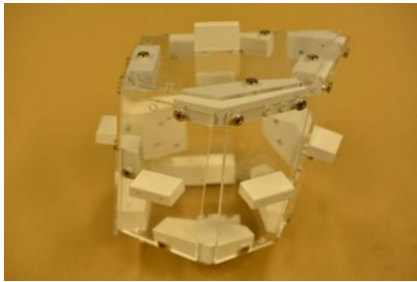
		
Task Variable	Low Complexity	High Complexity
1. Model (Norman, 1983; Pillay, 1997; Shalin et al., 1996)	Simple cubical box familiar to operator's mental representation	Abstract design unfamiliar to operator's mental representation
2. Symmetrical planes (Richardson et al., 2004, 2006)	High amount, for sides and cubical corner pieces, within and between parts	Low amount, for variably shaped sides and corner pieces, within and between parts
3. Components (Richardson et al., 2004, 2006)	Low amount	High amount
4. Component groups (Richardson et al., 2004, 2006)	Low amount (cf., 5 equally shaped and sized sides)	High amount (cf., all components unique in shape and size)
5. Selections (Richardson et al., 2004, 2006)	No redundant components and presented on work table in correct order	Redundant components and mixed presentation on work table
6. Fastenings (Richardson et al., 2004, 2006)	Only minimally required amount	High amount
7. Fastening points (Richardson et al., 2004, 2006)	No redundant holes on sides, and low amount of redundant holes and no faulty fastenings possible for corner pieces	High amount of redundant holes and irregular distribution over parts of redundant holes for sides and corner pieces
8. Novel (sub-)assemblies (Richardson et al., 2004, 2006)	Low amount	High amount, all sub-assemblies being different
9. Presenting orientation on work table (Pillay, 1997; Shalin et al., 1996)	Correct angle	All angled 90° randomly to the left or to the right along the horizontal plane.
10. Material (added by the authors)	Low transparency of acrylic glass sides	High transparency of acrylic glass sides, making visual perception more difficult

Table 1: Task variables manipulating assembly complexity.

2.4 Procedures

Under the guise of participating in a study that is 'exploring how people assemble', participants were told that eye-movements would be tracked. They were informed that there was no right or

wrong way for them to assemble and that they had all the time they needed to complete the assemblies.

The study took place in a quiet room with stable artificial lighting (550 lux). Participants were seated during the entire experiment in order to minimize differences in fatigue due to physical load. The work station (see Figure 1) consisted of a work table (A) on which all parts were displayed (B) and the assembly was to be made, and a table (C) placed at the opposite side for the presentation of the instructions, so the participant (D) consecutively had to turn around 180 degrees between visual intake of the instructions and execution of the assembly step. We artificially separated these two assembly phases for two reasons.

Firstly, this separation allowed to present the instructions automatically and in a standardized manner. This was important as pilot tests showed that when instructions would be presented at the work table, participants would be missing information because of switching back and forth between the presented parts and the presented instructions - again obstructing the manipulation to be the same for all participants. By running the instructions automatically, we could assure that the manipulation would be induced for a minimally same time window for all participants (contrary to operator-controlled instructions). Specifically, participants were asked to look at all the information presented on the screen of a laptop while not looking at the work table behind them, so that all participants were equally exposed to this part of the manipulation.

Secondly, by separating the instructions and the work table, unequal split-attention effects (cf., Sweller et al. 1998) were avoided, allowing us to ensure that a similar type of MWL was induced in both assemblies. The split-attention effect shows that procedural, compared to integrated instructions, induce more mental workload because of dividing attention between different information types (e.g., pictures and arrows vs. real-life objects, here, the assembly parts). Since the low complexity assembly is inherently easier to process cognitively, this condition would yield lower mental workload caused by split-attention as compared to the high complexity condition. This study however focused on manipulating complexity inherent to the assembly design (i.e., the intrinsic cognitive load caused by the element interactivity, Sweller et al. 1998) and excluded possible effects coming from extraneous cognitive load (Sweller, 2010) caused by, i.a., split-attention.

The instructions were presented on a 16 inch laptop (E) placed on a table, while participants were all seated at a normal distance (approximately 60 cm) from the screen. Both assemblies consisted

of 7 steps, so that both instructions presented 7 steps, each following the same format. Per step, pictures were shown (all equal in size, 1 per slide), along with very brief textual information on top of every slide (since we did not aimed to induce verbal processing). Each step first showed a picture of the two required parts placed next to the assembly-in-progress and subsequently, pictures of the part details. Then, a picture was shown for how to mount the two parts on the assembly-in-progress (indicated with arrows), a picture of the required result (with circles for where to screw) and pictures with details of this result. The average time windows of the instructions were comparable over conditions ($M_{low_complexity} = 43.17$ s; $M_{high_complexity} = 41.83$ s.).

After viewing the instructions for one assembly step, participants turned around and had to select the respective assembly parts needed out of an ordered display of parts for the low complexity assembly and a randomized display of parts (but fixed over participants) for the high complexity assembly (see Figure 2). Participants thus had to select, per step, two parts out of the parts displayed which also contained redundant parts in the high complexity condition. The two assemblies, laser cut in acrylic glass (PMMA; for the sides) and 3D-printed in polylactide (PLA; for the connection parts, handles and hinges) were to be screwed together with wood screws and a manual screw driver (F) (Pozidriv screw drive with magnetic tip).

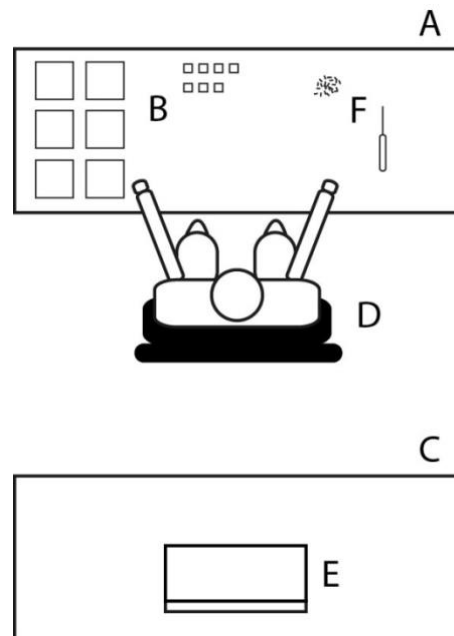


Figure 1: Top view on the work station, composed of the work table (A) with the assembly parts (B) and the screwdriver and screws (F), a table at the opposite side (C) with a laptop (E) displaying the instructions, and the seated participant operator (D).

Figure 2 summarizes the sequence of the assembly phases participants typically advanced through (based on Stork and Schubö 2010, see also Van Acker et al. 2018, for an assembly model case). First, the participant had to attend to the instructions by memorizing the two parts that were needed and how to orient them when mounting (i.e., stimulus preprocessing and mental rotation). When then facing the work table, visual inspection of the parts displayed on the table (i.e., feature extraction and stimulus identification) preceded final decision-making (i.e., executive control) on which parts to pick and how to orient them on the assembly-in-progress. Physical execution (action execution and motor adjustment) then followed, but iteratively fed back into the decision making process in which cognitive processes gauge whether the selected solution might indeed be the correct one. If not the case, decisions would be altered and physical execution would be performed over again. Because of the nature of complexity manipulation for real-life contexts, the high complexity condition would yield more of this trial and error, so that the iterative phase of decision making and physical execution would last longer, as compared to the low complexity condition. As participants might differ regarding their needed time for understanding and executing the prescribed instructions between and within steps, we compensated for these variations in our pupillometry analyses, by taking the mean pupil diameter per participant, per entire step (i.e., including instructions and execution).

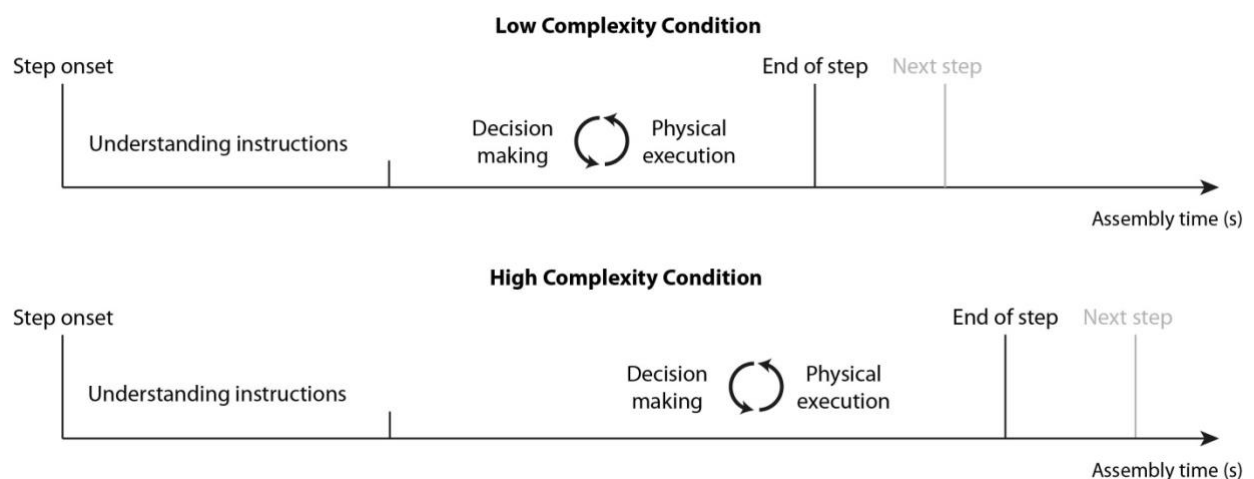


Figure 2: Sequence of the assembly phases of one step for the low and high complexity condition. Note that before Step 1 of each assembly, pictures of the finished assembly (i.e. the end goal) were shown.

2.5 Measures

In line with the suggestions by Van Acker et al. (2018) we assumed that our manipulation of the mental workload antecedent *assembly complexity* (see above), will interact with participants' cognitive architecture and that this task-human interaction will trigger firstly the spending of

cognitive resources as expressed by physiological reactions such as changes in *pupil size*. Indeed, by design, the element interactivity inherent to the assembly (cf., Sweller et al. 1998) will mainly trigger visual-spatial cognitive processing (i.e., mental rotation, object recognition, motor execution, etc., see Pillay 1997; Richardson et al. 2004, 2006) and working memory span processing (i.e., for keeping in mind the specific properties of the parts, e.g., amount of fastening points, see Stork and Schubö 2010) – all primarily related to the visual-spatial sketchpad part of working memory (Baddeley, 1992). As task complexity might also induce cognitive subjective experiences in participants according to Van Acker et al. (2018), we also measured this mental workload defining attribute through a *subjective estimation*.

Our resulting operational definition, necessary for solid theory building and comparability between empirical studies, stated that: "Mental workload will be reflected in a subjective estimation and in the pupil size, revealing an interplay between participants' limited working memory modalities on the one hand and the visual-spatial demands and working memory span demands being exposed to on the other".

2.5.1 Pupillometry

Pupil size was measured using the Eye Tracking Glasses (ETG) 2w of SensoMotoric Instruments (SMI; Teltow, Germany) at a binocular sampling rate of 60Hz. The SMI-ETG is a lightweight (47g) eye-tracking system with two small cameras integrated into the bottom rim of the frame of the glasses. The system deploys dark pupil tracking and automatic parallax compensation, and has an accuracy of 0.5° over all distances. Pupil size data were stored on a customized smartphone connected to the glasses with a cable and kept in the participants trouser pocket or waist bag. With respect to the analyses, the raw pupil signal was first extracted with the SMI BeGaze software. Next, blinks were linearly interpolated and the signal was filtered with a low-pass filter of 20Hz (note that we obtained similar findings with the more typical 10Hz filter). The start and end of each step in the assembly was used to define epochs for which the average pupil size was calculated. We report non-baselined data, but similar results were found when using a 5 second baseline that was extracted right before each step. We finally included a covariate in our statistical analyses controlling for potential pre-existing individual differences in resting-state pupil size (cf., Aminihaibashi, Hagen, Dyhre, Laeng, & Espeseth, 2019) by measuring participants' pupil size during the first two seconds in both the low and high complexity condition, to then calculate one overall average pupil size over both time periods/tasks. Figure 3 shows a picture of the SMI Eye Tracking Glasses. Data of the right pupil were used in all analyses, as common practice in pupillometry research.



Figure 3: The SMI-ETG worn during the experiment.

2.5.2 Subjective mental workload

The subjective experience of the mental workload was measured with a questionnaire averaging scores on three items to be rated on a 7-point Likert-scale (ranging from 1 *not at all* to 7 *to a great extent*; Cronbach's alpha = .82 for the low complexity condition and Cronbach's alpha = .86 for the high complexity condition).

Specifically, after completion of each assembly participants were asked to what extent they agreed to statements concerning the cognitive load experienced during the assembly process (i.e., 'I experienced this assembly as cognitively demanding') and the mental load (i.e., 'I had to invest mental effort while performing the assembly' and 'I experienced this assembly as mentally demanding'). These items were inspired on the work of cognitive load theory (F. G. Paas, 1992; Fred G. W. C. Paas, van Merriënboer, & Adam, 1994), the Subjective Workload Assessment Technique (SWAT; Reid and Nygren 1988) and the NASA-TLX (Hart & Staveland, 1988). We hereby covered the most prominent literature on subjective mental workload assessment (Rubio, Díaz, Martín, & Puente, 2004; Vidulich & Tsang, 1986), and accounted for the criticism on the NASA-TLX (de Winter, 2014).

2.5.3. Other measures

When simulating a more real-life work environment and thus letting go of a substantive amount of experimental control, applied contexts do allow for measuring variables affecting mental workload that cannot be experimentally controlled or are not feasible to control experimentally. First, participant's *visual-spatial intelligence* was measured, since this intelligence factor can largely affect interpersonal differences in assembly performance (cf., high spatial-ability learners devoting more cognitive resources in Mayer and Sims 1994). For this, we had all participants complete (a subset of) the Revised Minnesota Paper Form Board test (Stinissen, 1977) subsequent to the experiment on a different day (for practical reasons) and obtained a score per participant. Secondly, as we wanted to control for the potential effects of moderating variables outlined by Van Acker et al. (2018), we also measured *self-reported dexterity* (on a 7-point Likert-scale), *gender* and *age* (see Van Gerven et al. 2004, for the effects of aging on pupillary response under working memory load).

Thirdly, as mental workload can also be affected by *emotional load* and even *physical load* (Kahneman, 1973; Mandler, 1979; Norman & Bobrow, 1975), especially when measuring physiological reactions since they are thought to all draw from the same physiological resources, and since pupil size has also been found to be sensitive to emotional load (cf., Goldwater 1972; Wang et al. 2013), we also measured emotional and physical load. We subjectively gauged *physical load* with one item measuring to what extent one perceived the assembly as physically demanding and *emotional load* with five items - again with statements on the same 7-point Likert-scale. For the latter, we selected items from the Dundee Stress State Questionnaire (DSSQ; Matthews et al. 2013; Matthews 2016) pinpointing not only the negative emotions frustration and irritation, and whether participants felt tense during the assembly, but also by measuring the situation appraisal on coping ability and uncertainty potentially leading to these emotions. In so doing, we made a variable 'Emotional Load' with these items (Cronbach's $\alpha = .75$ for the low complexity condition and Cronbach's $\alpha = .91$ for the high complexity condition). We expected both scores to be low, indicated by a subjective rating below the scale midpoint (i.e., 4 on the 7-point Likert scale) and not to differ per condition.

Fourth, we checked for the work behavior variable *task engagement*, since the nature of the complexity manipulation might cause participants to easily get tired of the task. More precisely, the low complexity condition could lead to *mind wandering* and *being fed up with the task*, potentially associated with irritation and frustration, which importantly, might yield drops in assembly quality or higher levels of arousal gauged through pupillometry. A first item therefore asked (on the same a 7-point Likert-scale) to what extent the participant's mind started

wandering and the second item to what extent the participant was fed up with the task (again inspired on the DSSQ, Matthews et al. 2013; Matthews 2016). Since both items can indicate the amount of task engagement, but still measure different constructs (as shown by a low Cronbach's alpha estimation of internal consistency, being .66 for the low complexity condition and .62 for the high complexity condition), two separate variables 'Mind Wandering' and 'Fed Up' were included in the analyses.

Finally, as this is the first study applying pupillometry in an assembly context and as participants wore the eye-tracking glasses (ETG) in total for around 1 hour, we rated the device on 13 *wearability* items, using a 7-point Likert scale going from 1) not at all, to 7) completely. For this, we built on the Comfort Rating Scales of Knight et al. (2002), on Gemperle et al.'s (1998) design guidelines for wearability and Dunne and Smyth (2007) - all looking into the physical interaction of a user's body and the physical form of wearable devices and its effects on perceptions on the self and cognitive abilities. In so doing, we translated the parameters as proposed by these authors, being e.g., pressure and constriction, thermal balance, texture, moisture transport and freedom of movement, into items asking participants to what extent 1, they, after a while, forgot they still had the ETG on, 2. they felt that the ETG was moving, 3. caused irritating friction on the skin, 4. caused too much heat on the skin, 5. caused too much sweat on the skin, 6. caused too much pressure on the skin, 7. the ETG weighed too much, 8. caused pain, 9. obstructed movement, 10. they felt tense because of wearing the ETG, 11. the ETG caused distraction, and, 12. caused frustration.

To conclude the methods section: we have collected and made available all materials on the Open Science Framework - i.e., technical drawings of the assemblies, instructions, pictures of the display on the work table, questionnaires and the obtained subjective and physiological data. We did so in order to make future iterations on and replications of this approach as feasible as possible (see https://osf.io/uhmgv/?view_only=86c2843d9a0744c6a052c36cc503d43c).

3. Results and discussion

3.1 Task Complexity manipulation check

Instead of a repeated measures t-test, a Wilcoxon Signed Rank Test on perceived difficulty of the assemblies was selected because of violation of normality of the data. It confirmed that the high complexity condition was more difficult ($Md = 5.00$, range: 2-7, $N = 21$) to the participants as compared to the low complexity condition ($Md = 1.00$, range: 1-4, $N = 21$), $z = -4.04$, $p < .001$ (two-

tailed), with a large effect size of $r = .62$ (with $r = z/\sqrt{N_x + N_y}$, Rosenthal 1994). The complexity manipulation thus showed to be effective.

3.2 Emotional load, physical load, being fed up and mind wandering

In order to induce only mental workload, we wanted to minimize emotional and physical load processes as much as possible. Our emotional load (EL) variable showed sufficient reliability (see above), but lacked normality of data. We therefore first ran a Wilcoxon Signed Rank Test, revealing that EL was higher in the high complexity condition ($Md = 3.4$, range: 1.40-6.20, $N = 21$), as compared to the low complexity condition ($Md = 1.4$, range: 1-4.20, $N = 21$), $z = -3.83$, $p < .001$, with a large effect size of $r = .59$. Where in the low complexity condition Emotional Load was indeed perceived as almost non-existent, EL was thus however perceived as fairly low to average in the high complexity condition. The physiological data measured with pupillometry in the high complexity condition might thus not represent mental workload exclusively, but also to some extent emotional load - which is intelligible, since EL is hard to circumvent when facing complexity (cf., Van Acker et al. 2018) and pupil size indicates general arousal levels (Mathôt, 2018). Next, subjective ratings on physical load (PHL), showed to be equally low in both conditions ($Md = 1$, range_{low_complexity}: 1-2, range_{high_complexity}: 1-2, $N = 21$), not differing from each other, $z = -1.89$, $p = .06$.

Being Fed Up with the assembly, finally, resulted to be non-existent in the high complexity condition ($Md = 1$, range: 1-5, $N = 21$) but appeared not to differ from the scale midpoint ($Md = 4$, range: 1-6, $N = 21$) for the low complexity condition thus not meeting our expectation of being equally low for both conditions, $z = -2.88$, $p < .01$, with a medium effect size of $r = .59$. Mind Wandering was rated as low ($Md = 2$, range: 1-5, $N = 21$) in the high complexity condition and again appearing average to fairly high ($Md = 5$, range: 1-7, $N = 21$) for the low complexity condition, $z = -3.34$, $p < .01$, with a large effect size of $r = .52$.

3.3 Subjective mental workload experience

A Wilcoxon Signed Rank Test (selected because of violation of normality of the data) on our MWL variable also confirmed our manipulation, validated the subjective rating method, and indicates that participants perceived the mental workload to be higher in the high complexity condition ($Md = 4.67$, range: 1.67-7.00, $N = 21$), as compared to the low complexity condition ($Md = 1.67$, range: 1-4.33, $N = 21$), $z = -4.02$, $p < .001$, with a large effect size of $r = .62$. Hypothesis 1 was thus confirmed. All subjective measures are reported in Figure 4.

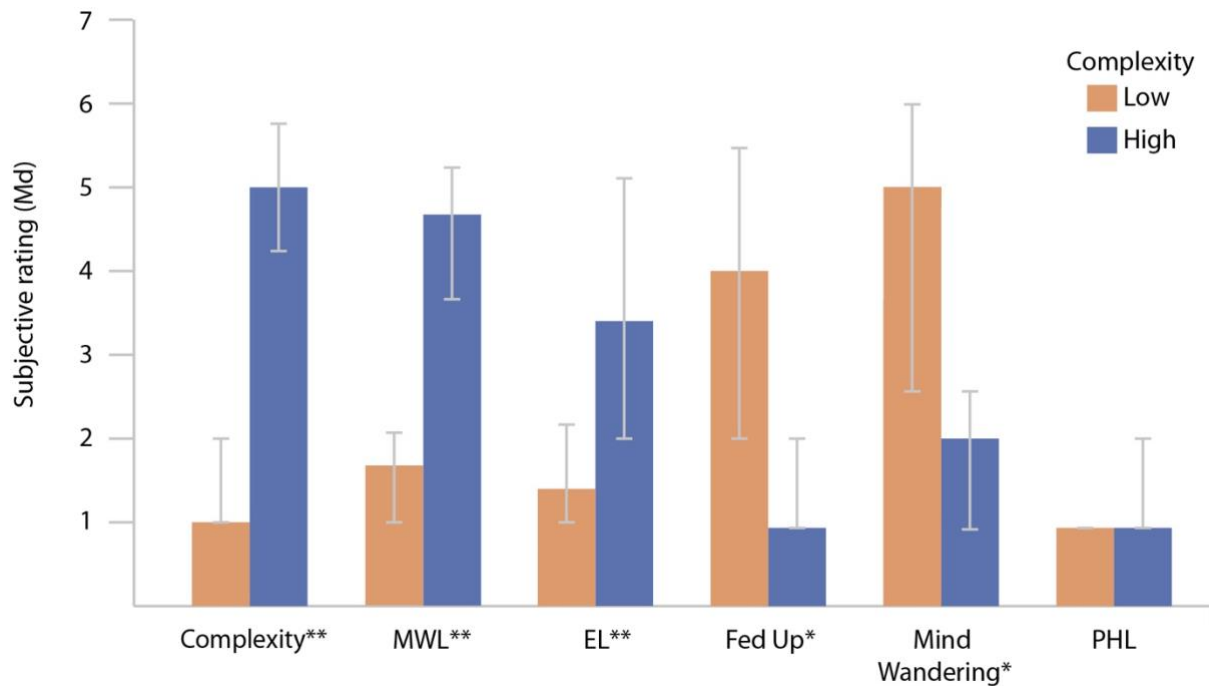


Figure 4: Overview of the medians of all subjective ratings, $N = 21$, $*p < .01$, $**p < .001$. Note: MWL = mental workload; EL = emotional load; PHL = physical load.

Due to the abnormality of the subjective scales and since $N = 21$ is not sufficient for correlational analyses, we refrain from exploring the effects of our moderator variables visual-spatial intelligence, dexterity, age and gender here.

3.4 Pupillometry

The data gathering of two participants was corrupted due to a technical malfunction. Two other participants dropped out before completion of the high complexity assembly. Other practical and technical impairments made that for some of the following analyses the sample sizes included were lower than foreseen, but still elaborate enough to run analyses.

3.4.1. Confirmatory analyses.

To statistically control for the confounding variables gender, age, dexterity, visual-spatial intelligence and resting-state pupil size, as discussed earlier, we first conducted a two-way repeated-measures ANCOVA with within-subjects factors Complexity and Step and these covariates of which the scores of the latter four were centered as recommended by Schneider et al. (2015). None of the covariates were overly correlated with one another (all $r_s < .53$). However, no covariate revealed effects adjusting the model to find significant differences, while age and visual-spatial intelligence interacted with the independent variable, thereby violating the homogeneity of regression slopes assumption for ANCOVA. We hence continued analyses

without statistical control and ran a two-way repeated-measures ANOVA with factors Complexity and Step. No main effect of Complexity was found, $F(1, 11) = 1.40$, $p = .26$ (partial eta squared = .11), nor was there a main effect of Step, $F(6,6) = 1.90$, $p = .09$ (partial eta squared = .15), or an interaction effect between Complexity and Step, $F(6,6) = .82$, $p = .56$ (partial eta squared = .07). More specifically, a paired samples t-test, showed that the mean pupil size over the entire Low Complexity condition was $M = 3.35$ ($SD = .44$, $N = 19$), not significantly differing from the mean pupil size over the entire High Complexity condition, being $M = 3.42$ ($SD = .61$, $N = 19$), $t = -.62$, $p = .54$, two-tailed, Cohen's $d = .14$. Hypotheses H2 and H3 were thus not supported. Figure 5 provides a plot of the two-way repeated measures ANOVA displaying the means and standard errors of all assembly steps compared. The error bars reveal the large amount of variation in pupil size.

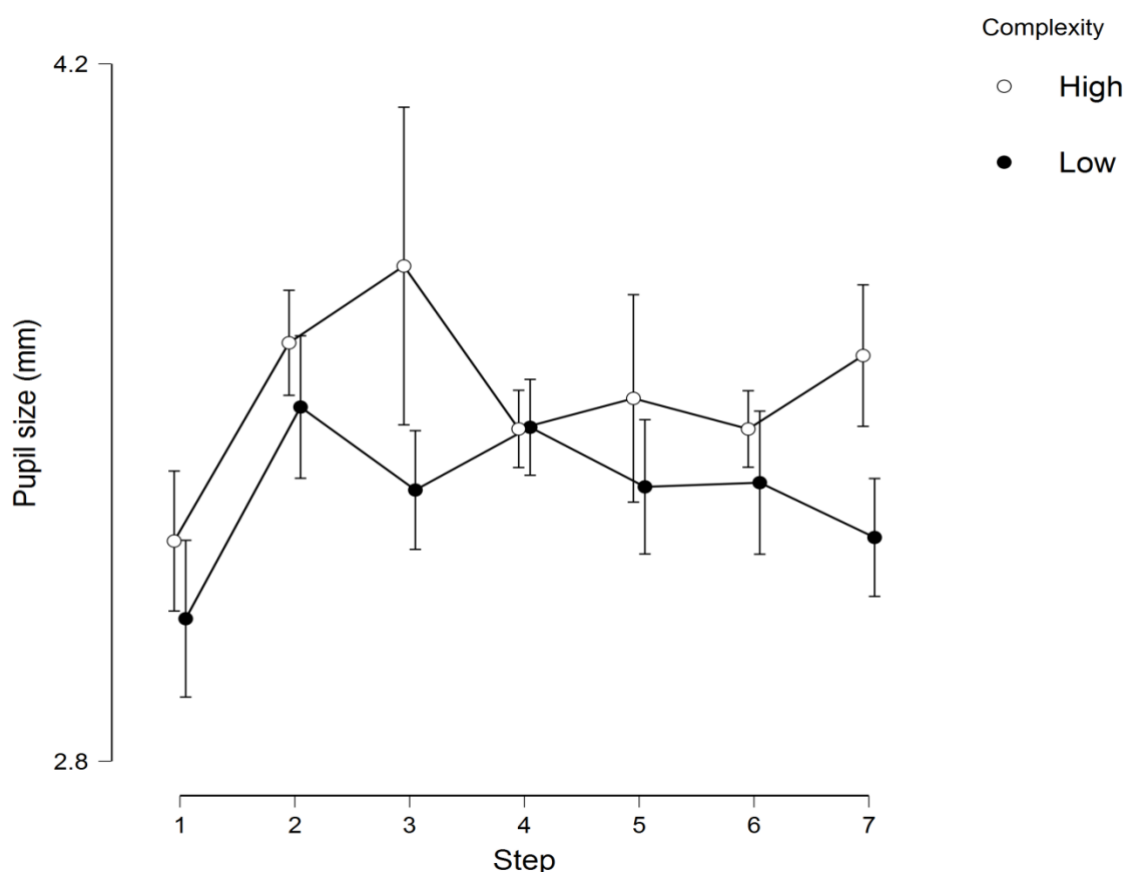


Figure 5: Mean pupil sizes and standard errors of the two-way repeated measures ANOVA for all steps in both conditions.

In all, our subjective measures all vouched for differences to be expected in arousal levels. Although our pupillometry hypotheses were not backed by significant differences in the data. For these confirmatory analyses addressing our hypotheses, the entire sequence of assembly

execution as presented in Figure 2 was covered. In the following exploratory set of analyses (i.e., as not hypothesized) we wanted to exclude possible noise coming from, e.g., movement and random light reflection effects on the work table, by restricting to the first 30s of each assembly step. During these time windows participants merely looked at a screen presenting pictures of the end goal and instructions of each step.

3.4.2. Exploratory analyses.

Scrutinizing for possible effects of the first 30s of all steps and now also of viewing pictures of the end results right before the first step, also here, we first conducted a two-way repeated-measures ANCOVA with factors Complexity and Step and added the covariates gender, age, dexterity, visual-spatial intelligence and resting-state pupil size. No covariate revealed effects adjusting the model to find significant differences. Resting-state pupil size moreover interacted with Complexity thereby violating the homogeneity of regression slopes assumption for ANCOVA. Next, a two-way repeated-measures ANOVA with factors Complexity and Step showed no main effect of Complexity, $F(1, 14) = 3.08, p = .10$, partial eta squared = .18, but did reveal a main effect for Step, $F(2.76, 38.57) = 10.54, p < .001$, partial eta squared = .43 (consulting a Greenhouse-Geisser estimates of sphericity, $\epsilon = .39$, correcting the degrees of freedom after violation of the assumption of sphericity as shown by Mauchly's Test of Sphericity, $\chi^2(27) = 61.91, p < .001$) and did not display an interaction effect between Complexity and Step, $F(3.51, 49.15) = 1.59, p = .20$, partial eta squared = .10 (consulting a Greenhouse-Geisser estimates of sphericity, $\epsilon = .50$, correcting the degrees of freedom after violation of the assumption of sphericity as shown by Mauchly's Test of Sphericity, $\chi^2(27) = 58.63, p < .01$).

Elaborating on the main effect of Steps, a paired samples t-test showed that viewing the first 30s of the end goal and the first 30s of the instructions in the final step as well induced a significantly larger pupil size in the High Complexity condition compared to the Low Complexity condition ($M = 3.25$ vs. $M = 3.04$ and $M = 3.69$ vs. $M = 3.35, ps < .05$, with Cohen's d effect sizes of .54 and .69, respectively). The other steps did not reveal significant differences (see Table 2 for all results).

	Low Complexity		High Complexity		<i>t</i>	<i>df</i>	<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
End goal view	3.04	.42	3.25	.64	-2.36	18	<u>.03</u>
Step 1	2.94	.39	3.05	.61	-1.08	17	.30
Step 2	3.26	.45	3.30	.53	-.56	17	.59
Step 3	3.33	.48	3.41	.63	-.74	18	.47
Step 4	3.28	.46	3.38	.58	-1.30	18	.21
Step 5	3.32	.49	3.35	.67	-.26	16	.80
Step 6	3.21	.52	3.40	.61	-1.78	15	.10
Step 7	3.35	.45	3.69	.74	-2.69	14	<u>.02</u>

Table 2: Paired samples *t*-test results comparing pupil dilation (in mm) in the Low vs. High Complexity condition for the first 30s of all steps and viewing the end goal.

To look into variations between steps *within* conditions, we first ran a one-way repeated measures ANOVA on Low Complexity ($N = 16$). Mauchly's Test of Sphericity also here showed violation of the assumption of sphericity, $\chi^2(27) = 51.55$, $p < .01$. Greenhouse-Geisser estimates of sphericity ($\epsilon = .51$) corrected the degrees of freedom. Results this time indicated a significant difference between the steps within this condition, $F(3.58, 53.57) = 6.28$, $p < .01$, with a large effect size as shown by a partial eta squared value of .30. Post hoc tests using Bonferroni correction revealed that viewing the end goal ($M = 3.04$, $SD = .46$) induced a significantly smaller pupil than step 7 ($M = 3.34$, $SD = .44$), $p < .05$. Then, the first step ($M = 2.92$, $SD = .41$) also appeared to trigger a smaller pupil size than step 3 ($M = 3.28$, $SD = .50$), $p < .05$, and compared to step 4 ($M = 3.29$, $SD = .50$), 5 ($M = 3.27$, $SD = .48$), 6 ($M = 3.17$, $SD = .49$) and 7 ($M = 3.34$, $SD = .44$), all $ps < .01$.

Finally, for the High Complexity condition ($N = 17$), Mauchly's Test of Sphericity showed violation of the assumption of sphericity, $\chi^2(27) = 76.70$, $p < .001$. Greenhouse-Geisser estimates of sphericity ($\epsilon = .46$) corrected the degrees of freedom. Results showed a significant difference between the steps within this condition as well, $F(3.22, 51.48) = 5.22$, $p < .01$, with a large effect size as indicated by a partial eta squared value of .25. Here, post hoc tests using Bonferroni correction showed that the first step ($M = 3.07$, $SD = .62$) again induced a significantly smaller pupil than step 7 ($M = 3.65$, $SD = .71$), $p < .01$. Remarkably, the pupil size during the last step was

the largest of all steps, while this step was the easiest of all since only one component was left on the work table.^{1,2}

Concluding, the effect that viewing the end goal and the final step triggered a larger pupil size in the High Complexity condition tallies with what we expected from the complexity manipulation, but the observation that the other steps did not show significant differences (cf., H3) and that both final steps induced the highest pupil sizes was unexpected. In hindsight and tentatively, this latter result could fit with the idea that cognitive resources - being partly physiological in nature - are limited and hence can become depleted over time, leading to cognitive fatigue (cf., Hockey 1997). We however consider measurement error also for the exploratory analyses (see Table 2) too vast to draw conclusions.

3.5 Wearability

Finally, descriptive analyses showed that the ETG device was perceived as fairly wearable in terms of feeling the device move ($M = 2.62$), causing irritating friction ($M = 2.24$), causing too much heat on the skin ($M = 2.38$), causing too much sweat on the skin ($M = 1.90$), weighing too much ($M = 2.81$), obstructing movement ($M = 2.70$), feeling tense because of wearing the device ($M = 1.81$), causing distraction ($M = 2.48$) and causing frustration ($M = 1.88$). The device did however seem to cause too much pressure on the skin ($M = 3.81$) and some pain ($M = 3.29$), while it scored low when asked if the participant had forgotten to be wearing the device ($M = 2.62$). Answers to open-ended questions suggest that the pain was primarily caused by the frame around the right ear and the pressure by the nose bridge part of the frame, pointing at the need for specific product redesign iterations.

4. General discussion

The current pilot study built upon prior laboratory (e.g., Kahneman et al. 1969) and applied (e.g., Truschinski et al. 2018) pupillometry research and was, to our knowledge, the first to apply mobile pupillometry in a manual assembly task. We designed a procedure in which assembly complexity was systematically manipulated based on assembly complexity literature (Richardson et al., 2004, 2006), intrinsic cognitive load was isolated based on cognitive load theory (Sweller,

¹ After controlling for our (centered) covariates with a repeated measures ANCOVA, only gender showed to have an effect, $p < .05$ for both conditions. Since eventually only 33.3% of the total sample, $N = 19$, were female, we withdraw from making inferences about this covariate.

² Note that we then also ran the same tests within conditions on the mean pupil sizes for the entire steps (cf., the confirmatory analyses). After correcting the degrees of freedom, no significant differences between the steps within Low Complexity were found, $F(2.51, 30.12) = 1.75$, $p = .19$ (partial eta squared = .13), nor within High Complexity, $F(2.28, 31.86) = 1.25$, $p = .30$ (partial eta squared = .08).

1994, 1988) and confounding variables present in real-life contexts were methodologically accounted for based on the overarching mental workload framework of Van Acker et al (2018). Being explorative in nature, this pilot study also served to validate the design of two real-life manual assemblies of diverging complexity while allowing to pinpoint the wearability of mobile pupillometry.

The complexity manipulation showed to be effective as participants experienced the high complexity assembly to be more difficult and reported a higher subjective MWL compared to the low complexity condition, thereby confirming our first hypothesis (H1) and conceptually replicating Richardson et al. (2004, 2006). However, some degree of unanticipated emotional load was observed in the high complexity condition, while the low complexity condition was characterized by moderate levels of being fed up with the assembly and mind wandering. The former might have caused the pupil dilation data observed to include negative emotional variance next to MWL variance. Being fed up with the assembly and mind wandering in the low complexity condition did not result in reported negative EL load. Experimentally manipulating MWL in the field hence does raise some challenges and also unexpected negative EL or decreased task engagement affecting MWL-measurement could be hard to avoid. Future research could do this more profoundly in order to grasp the measure's selectivity (Gerald Matthews, Reinerman-Jones, Barber, et al., 2015) with, e.g., emotion recognition technology and user centered design methodologies (e.g., Jokinen 2015) not only helping in fine-grained MWL-optimization, but more general workload-optimization (including emotional load, task engagement, physical load, etc.).

Despite the strong statistical effects for the subjective ratings (H1), the external validity of the pupillometry as covered by our hypotheses H2 and H3, was not confirmed. Our exploratory analyses unveiled non-conclusive significant effects. Specifically, viewing the end goal of the high complexity assembly induced larger pupil sizes when compared to the low complexity condition. So did viewing the instructions of the last, easiest step as this step triggered the largest pupil sizes of all steps. This latter observation might, tentatively, be explained through cognitive fatigue (cf., Hockey 1997), reflecting the depletion of participants' cognitive resources over time. Still, in the case of these exploratory analyses, we coin these observations as non-conclusive, since all other assembly steps did not reveal significant effects and also here, measurement error was substantial.

These findings together suggest that deploying pupillometry on the shop floor might not yet be feasible when following the current laboratory-based procedures as we did. Aiming to foster the

development of field-sensitive measurement protocols regarding pupillometry, we formulate below five speculative explanations for the lack of measurement sensitivity we observed.

First, the natural variation in luminance coming from the parts of the assembly, the work table, screws, etc., varied per momentary participant assembly strategy and might have thwarted the signal-to-noise ratio. As shown by Palinko and Kun (2012a), the pupil not only reacts to ambient light, but also to luminance of the light around the fixation point, so that even small visual targets in a driving simulator (with an angular radius of 2.5°) showed to affect pupil size. These researchers refer to driving contexts in which participants scan visual targets of different size and luminance when attending to traffic (e.g. vehicles, traffic signs). The assembly parts, tools, work table and instruction pictures that we used also varied in luminance and so do most objects in industrial environments. Purely visually attending to objects might thus have caused pupil sizes to be more strongly affected by luminance than by MWL, despite of stable ambient artificial light. A future solution to this could be to first measure and then subtract the mere light reaction from the task-related pupil signal in order to isolate the MWL-related pupil size (Kun, Palinko, & Razumenić, 2012; O Palinko & Kun, 2012; Oskar Palinko & Kun, 2012; Oskar Palinko, Kun, Shyrokov, & Heeman, 2010), to use machine learning (Haar-like) feature selection (Wang et al., 2013), leverage the ratio between low and high frequency bands as in Peysakhovich et al. (2015) or to build on Duchowski et al.'s (2018) Index of Pupillary Activity. Note that eventually also the constant (re-)focusing of the eye (cf., Mathôt 2018) when attending to a real-life task (instead of fixating on a screen at the same distance in laboratory studies) neither helps in attaining a feasible signal-to-noise ratio.

A second possible explanation why our hypotheses on the pupil observations were not confirmed, might be rooted in the curvilinear relation between task difficulty and pupil size. Granholm et al. (1996), intriguingly, noted that higher MWL induces a greater pupil diameter, but only up to a point where mental overload causes the pupil size to decrease again. Since we also observed moderate levels of negative EL (possibly caused by a negative coping appraisal, as measured by one of the EL-scale items), cognitive overload might actually have been unintentionally induced. Pupil sizes may hence have recurrently decreased when attending to the high complexity condition. Future research is thus needed to explore how laboratory findings on a curvilinear trend translate to field settings. Again, such notions stress the importance of systematically including EL into cognitive load measurement, since EL could be one indicator of cognitive overload.

Third, we could have done better at elaborating on what resulted from participants being fed up and starting mind wandering when performing the low complexity assembly. One not so far-fetched inference might be that they engaged in more cognitive processing related to task-unrelated personal thoughts. This notion might be backed by Franklin et al.'s (2013) seemingly contradictory finding that spontaneous mind-wandering in reading tasks induced higher tonic pupil dilation (also see, Algermissen, 2019). The reason for this would be that the locus coeruleus-norepinephrine system, a neurotransmitter system involved in modulating arousal and attention and related to pupil dilation, makes people to explore for alternative tasks when the current task is perceived as non-relevant (Smallwood et al., 2012), that is, when disengagement from the task occurs (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010). This 'exploration phase' has been related to large pupils, while the opposite 'exploitation phase' is accompanied by intermediate pupil sizes (Mathôt, 2018). A more profound user-centered approach trying to unveil the underlying cognitive and emotional processing while attending to a task hence appears highly desirable again, in order to fully understand the physiological outcomes.

Fourth, in the current experiment, participants exerted physical effort when mounting and screwing parts together. Interestingly, Jiang et al. (2015) discovered that pupil size increases when visual-motor aiming complexity increases. Since our high complexity condition inferred more visual-motor aiming execution, this effect on pupil diameter should have in fact confounded our observations in favor of our hypotheses. The story becomes more complex however, when taking into account Fletcher et al.'s (2017) finding that precise motor movement contrarily causes pupil constriction during response preparation and execution. Our exploratory analyses excluded movement by using a 30s time window and did yield significant differences, possibly pointing at the influence of movement. These effects are however too random to draw conclusions from at this stage of mobile pupillometry research. Future work could therefore introduce movement capturing technology such as an accelerometer into the measurement protocol and, for instance, prior to MWL-measurement map to what extent the specific movements required in a specific task affect pupil size. This way and in the long term, machine learning algorithms could also help in calibrating MWL-estimates based on movement artifacts, next to the effects of light.

Finally, the deployed statistical power leveraged by a sample size of $N = 19$ might be too limited to support solid pupillometry in such a noisy environment - as compared to controlled laboratory environments. Some results suggests this worth considering given that the main effect of the Steps for the confirmatory analyses was minimally significant ($p = .09$), while also the main effects of Complexity ($p = .10$) and Step ($p < .001$) for the exploratory analyses could suggest that a larger

sample size could have yielded a higher signal-to-noise ratio, potentially revealing more significant differences. Especially in the light of the missing data as we encountered here (and its consequences for the ANOVA's when comparing steps, e.g., $N = 12$, $N = 16$) future research could try to upscale similar experimental procedures or use a Bayesian sequential approach when collecting data (cf., Nathoo and Masson 2016; Etz and Vandekerckhove 2018). Still, the economic and practical restrictions of work environments will, in the long run, require sample sizes to be substantially smaller than $N = 19$. With our pilot sized sample, we showed that the above suggestions will also have to be addressed before mobile pupillometry could become valid and reliable for small sampled trials.

In terms of wearability, the eye-tracking glasses showed to be fairly wearable in terms of mental comfort, distraction and physical properties. Some re-design iterations are still required for the type of eye-tracking glasses we used, in order not to cause pain and pressure for all people. Also the awareness of having the glasses on was prevalent in our setting, posing a second focus for future redesign. In all, these data do suggest that mobile pupillometry deploying eye-tracking glasses could be feasible in the future in terms of wearability. Interesting here would be to now test wearability during a longer time span, in even more mobile situations and to include user acceptance measures related to privacy concerns, for example. Also comparison with other objective MWL-measures such as EEG and fNIRS would help in differentiating which measures are best suited for specific industrial contexts in relation to operator tasks, time windows of interest, safety regulations (cf., wearing helmets or safety glasses), etc.

In all, our results could not replicate seminal laboratory work on pupillometry (cf., Beatty & Kahneman, 1966; Hess & Polt, 1960; Kahneman et al., 1969; or more recently, Krejtz, Duchowski, Niedzielska, Biele, & Krejtz, 2018; Moyes, Sari-sarraf, & Gilbert, 2019) and highlights the challenges in reaching external validity. Research on driving (cf., Recarte & Nunes, 2003), air traffic control (cf., Ahlstrom & Friedman-berg, 2006) and, for instance, piloting (cf., Causse et al., 2016) and surgery performance (cf., Dalveren et al., 2018; Erridge et al., 2017) already made great strides in closing this gap. While being seated and at a predominantly stable distance to experimental the stimuli, pupillometry in these contexts might become robust as soon as, i.a., variations in luminance (Lohani, Payne, & Strayer, 2019) and curvilinear effects can be accounted for. Pupillometry validation in assembly work will now need to step into this latter direction too, but confounding of physical effort or freedom of head movement will show to be equally important, so we foresee. These diverse fields in ergonomics and human factors could strongly benefit from

joining forces, in that respect, and learn from the different boundary conditions they are confronted with.

Our study in specific took up the challenge of exploring the wearability and external validity of a robust laboratory measure integrated into a wearable device, and thereby taps into the needs of the assembly workplace. Before user acceptance of mobile pupillometry by companies, but especially by operators can be achieved (Van Acker, Conradie, Vlerick, & Saldien, 2019), important steps will have to be made and we hope the above-mentioned findings and reflections could help in doing so. Additionally, the authors were strongly motivated to address the recent call for field research (by, e.g., Maner, 2016), extending the veracity and robustness of controlled psychological laboratory experiments in which stimuli are typically simplified and participants isolated from the real world (Neisser, 1976). The outcomes of the presented work show how challenging field applications of wearable physiological MWL-measurement in specific can be and therefore invite a more nuanced view on the current state of science and its implementability. More in-depth knowledge on contextual factors intricately affecting measurement validity is therefore imperative for future wearable MWL-measurement development. We hope that the implementation of confounding variables into measurement interpretation as performed here could help stimulate the development of more elaborate contextual measurement procedures.

5. Conclusion

Objective wearable mental workload measures will be imperative in Industry 4.0. This pilot study set out to explore the wearability and external validity of pupillometry, a physiological indicator of mental workload. Participants performed an assembly of low and one of high complexity while wearing eye-tracking glasses. Overall, the latter were perceived as fairly wearable. While subjective mental workload did differ significantly, mean pupil size did not differ between both assemblies. Current pupillometry procedures might thus not be suited for in-the-field mobile pupillometry yet. Some suggestions on possible future directions were therefore proposed. In all, we strongly look forward to seeing iterations on this pilot endeavor and hope we have provided crucial insights to eventually make real-world mobile pupillometry a MWL-measure of the future.

6. Funding

This work was supported by the strategic research centre for the manufacturing industry Flanders Make, Oude Diestersebaan, 133, 3920 Lommel, Belgium, as part of the SBO project 'Augmented workers using smart robots in a manufacturing cell (Yves)'.

7. References

- Ahlstrom, U., & Friedman-berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, 36, 623–636. <https://doi.org/10.1016/j.ergon.2006.04.002>
- Algermissen, J., Bijleveld, E., Jostmann, N. B., & Holland, R. W. (2019). Explore or reset? Pupil diameter transiently increases in self-chosen switches between cognitive labor and leisure in either direction. *Cognitive, Affective, & Behavioral Neuroscience*, x(x), x–x. <https://doi.org/10.3758/s13415-019-00727-x>
- Aminihajibashi, S., Hagen, T., Dyhre, M., Laeng, B., & Espeseth, T. (2019). Individual differences in resting-state pupil size: Evidence for association between working memory capacity and pupil size variability. *International Journal of Psychophysiology*, 140(March), 1–7. <https://doi.org/10.1016/j.ijpsycho.2019.03.007>
- Annett, J. (2002). A note on the validity and reliability of ergonomics methods. *Theoretical Issues in Ergonomics Science*, 3(2), 228–232. <https://doi.org/10.1080/14639220210124067>
- Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4), 425–438. <https://doi.org/10.1007/s10648-010-9130-y>
- Atman, C. J., Adams, R. S., Cardella, M. F., Turns, J., Mosborg, S., & Saleem, J. (2007). Engineering design processes: A comparison of students and expert practitioners. *Journal of Engineering Education*, 96(4), 359–379. <https://doi.org/10.1002/j.2168-9830.2007.tb00945.x>
- Ayaz, H., Onaral, B., Izzetoglu, K., Shewokis, P. A., McKendrick, R., & Parasuraman, R. (2013). Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: Empirical examples and a technological development. *Frontiers in Human Neuroscience*, 7(871), 1–13. <https://doi.org/10.3389/fnhum.2013.00871>
- Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, 59(1), 36–47. <https://doi.org/10.1016/j.neuroimage.2011.06.023>
- Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556–559. <https://doi.org/10.1126/science.1736359>
- Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, 5(10), 371–372. <https://doi.org/10.3758/BF03328444>
- Bombeke, K., Duthoo, W., Mueller, S. C., Hopf, J. M., & Boehler, C. N. (2016). Pupil size directly modulates the feedforward response in human primary visual cortex independently of attention. *NeuroImage*, 127, 67–73. <https://doi.org/10.1016/j.neuroimage.2015.11.072>
- Boucsein, W. (2012). *Electrodermal activity* (2nd ed.). New York, NY: Springer.
- Braem, S., Coenen, E., & Bombeke, K. (2015). Open your eyes for prediction errors. *Cognitive, Affective and Behavioral Neuroscience*, 15, 374–380. <https://doi.org/10.3758/s13415-014-0333-4>
- Brolin, A., Thorvald, P., & Case, K. (2017). Experimental study of cognitive aspects affecting human performance in manual assembly. *Production and Manufacturing Research*, 5(1), 141–163. <https://doi.org/10.1080/21693277.2017.1374893>
- Brookhuis, K. A., & De Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics*, 36(9), 1099–1110. <https://doi.org/10.1080/00140139308967981>
- Causse, M., Peysakhovich, V., & Fabre, E. F. (2016). High working memory load impairs language processing during a simulated piloting task: An ERP and pupillometry study. *Frontiers in Human Neuroscience*, 10(240), 1–14. <https://doi.org/10.3389/fnhum.2016.00240>
- Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, 74, 221–232. <https://doi.org/10.1016/j.apergo.2018.08.028>
- Collet, C., Salvia, E., & Petit-Boulanger, C. (2014). Measuring workload with electrodermal activity during common braking actions. *Ergonomics*, 57(6), 886–896. <https://doi.org/10.1080/00140139.2014.899627>
- Dalveren, G. G. M., Cagiltay, N. E., Ozcelik, E., & Maras, H. (2018). Mental workload of surgical

- residents: feasibility of an educational computer-based surgical simulation environment (ECE) considering the hand condition. *Surgical Innovation*, 25(6), 616–624.
<https://doi.org/10.1177/1553350618800078>
- de Winter, J. C. F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. *Cognition, Technology and Work*, 16(3), 289–297.
<https://doi.org/10.1007/s10111-014-0275-1>
- Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., ... Giannopoulos, I. (2018). The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–13. <https://doi.org/10.1145/3173574.3173856>
- Dunne, L. E., & Smyth, B. (2007). Psychophysical elements of wearability. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 299.
<https://doi.org/10.1145/1240624.1240674>
- Erridge, S., Ashraf, H., Purkayastha, S., Darzi, A., & Sodergren, M. H. (2017). Comparison of gaze behaviour of trainee and experienced surgeons during laparoscopic gastric bypass. In *12th Annual Academic Surgical Congress* (pp. 287–294). Las Vegas, Nevada, USA.
<https://doi.org/10.1002/bjs.10672>
- Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin and Review*, 25(1), 5–34. <https://doi.org/10.3758/s13423-017-1262-3>
- Fletcher, K., Neal, A., & Yeo, G. (2017). The effect of motor task precision on pupil diameter. *Applied Ergonomics*, 65, 309–315. <https://doi.org/10.1016/j.apergo.2017.07.010>
- Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, 73(April), 90–99.
<https://doi.org/10.1016/j.apergo.2018.06.006>
- Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *Quarterly Journal of Experimental Psychology*, 66(12), 2289–2294. <https://doi.org/10.1080/17470218.2013.858170>
- García, A., David, C., Vera, J., & Jim, R. (2017). Intraocular pressure is sensitive to cumulative and instantaneous mental workload, 60, 313–319. <https://doi.org/10.1016/j.apergo.2016.12.011>
- Gemperle, F., Kasabach, C., Stivoric, J., Bauer, M., & Martin, R. (1998). Design for wearability. In *ISWC '98 Proceedings of the 2nd IEEE International Symposium on Wearable Computers* (p. 116). Washington, D.C. Retrieved from <http://dl.acm.org/citation.cfm?id=857199.857998>
- Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective and Behavioral Neuroscience*, 10(2), 252–269. <https://doi.org/10.3758/CABN.10.2.252>
- Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin*, 17(5), 340–355. <https://doi.org/http://dx.doi.org/10.1037/h0032456>
- Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, 3, 457–461.
<https://doi.org/https://doi.org/10.1111/j.1469-8986.1996.tb01071.x>
- Guru, K. A., Shafiei, S. B., Khan, A., Hussein, A. A., Sharif, M., & Esfahani, E. T. (2015). Understanding cognitive performance during robot-assisted surgery. *Urology*, 86(4), 751–757. <https://doi.org/10.1016/j.urology.2015.07.028>
- Hairston, W. D., Whitaker, K. W., Ries, A. J., Vettel, J. M., Bradford, J. C., Kerick, S. E., & McDowell, K. (2014). Usability of four commercially-oriented EEG systems. *Journal of Neural Engineering*, 11(4). <https://doi.org/10.1088/1741-2560/11/4/046018>
- Hansen, J. P., Mardanbegi, D., Biermann, F., & Bækgaard, P. (2018). A gaze interactive assembly instruction with pupillometric recording. *Behavior Research Methods*, 50(4), 1723–1733.
<https://doi.org/10.3758/s13428-018-1074-z>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload*. (pp. 139–183). Amsterdam: Elsevier Science Publishers B.V. (North-Holland).

- Heine, T., Lenis, G., Reichensperger, P., Beran, T., Doessel, O., & Deml, B. (2017). Electrocardiographic features for the measurement of drivers' mental workload. *Applied Ergonomics*, 61, 31–43. <https://doi.org/10.1016/j.apergo.2016.12.015>
- Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, 132, 349–350.
- Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: a cognitive energetical framework. *Biological Psychology*, 45(96), 73–93. [https://doi.org/10.1016/S0301-0511\(96\)05223-4](https://doi.org/10.1016/S0301-0511(96)05223-4)
- Hoedt, S., Claeys, A., Van Landeghem, H., & Cottyn, J. (2017). The evaluation of an elementary virtual training system for manual assembly. *International Journal of Production Research*, 7543, 1–13. <https://doi.org/10.1080/00207543.2017.1374572>
- Howard, S. J., Burianová, H., Ehrich, J., Kervin, L., Calleia, A., Barkus, E., ... Humphry, S. (2015). Behavioral and fMRI evidence of the differing cognitive load of domain-specific assessments. *Neuroscience*, 297, 38–46. <https://doi.org/10.1016/j.neuroscience.2015.03.047>
- Huang, L. Y., She, H. C., Chou, W. C., Chuang, M. H., Duann, J. R., & Jung, T. P. (2013). Brain oscillation and connectivity during a chemistry visual working memory task. *International Journal of Psychophysiology*, 90(2), 172–179. <https://doi.org/10.1016/j.ijpsycho.2013.07.001>
- Jiang, X., Zheng, B., Bednarik, R., & Atkins, M. S. (2015). Pupil responses to continuous aiming movements. *International Journal of Human Computer Studies*, 83, 1–11. <https://doi.org/10.1016/j.ijhcs.2015.05.006>
- Jokinen, J. P. P. (2015). Emotional user experience: Traits, events, and states. *International Journal of Human Computer Studies*, 76, 67–77. <https://doi.org/10.1016/j.ijhcs.2014.12.006>
- Kahneman, D. (1973). Attention and effort. *The American Journal of Psychology*, 88(2), 339. <https://doi.org/10.2307/1421603>
- Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*, 79(1), 164–167. <https://doi.org/10.1037/h0026952>
- Katidioti, I., Borst, J. P., Haan, D. J. B. De, Pepping, T., Vugt, M. K. Van, & Taatgen, N. A. (2016). Interrupted by your pupil: An interruption management system based on pupil dilation. *International Journal of Human-Computer Interaction*, 32(10), 791–801. <https://doi.org/10.1080/10447318.2016.1198525>
- Knight, J., Baber, C., Schwartz, A., & Brostow, H. (2002). The comfort assesment of wearable computers. In *6th International Symposium on Wearable Computers (ISWC'02)*.
- Kocielnik, R., Sidorova, N., Maggi, F. M., Ouwerkerk, M., & Westerink, J. H. D. M. (2013). Smart technologies for long-term stress monitoring at work. *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems*, (June), 53–58. <https://doi.org/10.1109/CBMS.2013.6627764>
- Kolbeinsson, A., Thorvald, P., & Lindblom, J. (2017). Coordinating the interruption of assembly workers in manufacturing. *Applied Ergonomics*, 58, 361–371. <https://doi.org/10.1016/j.apergo.2016.07.015>
- Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS ONE*, 13(9), 1–23 e0203629. <https://doi.org/10.1371/journal.pone.0203629>
- Kun, A. L., Palinko, O., & Razumenić, I. (2012). Exploring the effects of size and luminance of visual targets on the pupillary light reflex. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '12* (pp. 183–186).
- Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing*, 12(1), 13–21. <https://doi.org/10.1007/s10339-010-0370-z>
- Lohani, M., Payne, B. R., & Strayer, D. L. (2019). A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in Human Neuroscience*, 13(57), 1–27. <https://doi.org/10.3389/fnhum.2019.00057>

- Longo, F., Nicoletti, L., & Padovano, A. (2017). Smart operators in industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context. *Computers and Industrial Engineering*, 113, 144–159. <https://doi.org/10.1016/j.cie.2017.09.016>
- Mandler, G. (1979). Thought processes, consciousness, and stress. In V. Hamilton & D. M. Warburton (Eds.), *Human stress and cognition: an information processing approach*. (pp. 179–201). New York: John Wiley & Sons, Inc.
- Maner, J. K. (2016). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*, 66, 100–106. <https://doi.org/10.1016/j.jesp.2015.09.018>
- Marinescu, A. C., Sharples, S., Ritchie, A. C., Sánchez López, T., McDowell, M., & Morvan, H. P. (2018). Physiological parameter response to variation of mental workload. *Human Factors*, 60(1), 31–56. <https://doi.org/10.1177/0018720817733101>
- Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(16), 1–23. <https://doi.org/10.5334/joc.18>
- Matthews, G. (2016). Multidimensional profiling of task stress states for human factors: A brief review. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(6), 801–813. <https://doi.org/10.1177/0018720816653688>
- Matthews, Gerald, Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors*, 57(1), 125–143. <https://doi.org/10.1177/0018720814539505>
- Matthews, Gerald, Reinerman-Jones, L., Wohleber, R., Lin, J., Mercado, J., & Abich, J. (2015). Workload is multidimensional, not unitary: what now? In D. D. Schmorow & C. M. Fidopiastis (Eds.), *Foundations of augmented cognition: 9th International Conference, AC 2015, held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings* (pp. 44–55). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20816-9_5
- Matthews, Gerald, Szalma, J., Rose, A., Neubauer, C., & Warm, J. S. (2013). Profiling task stress with the dundee stress state questionnaire. In L. Cavalcanti & S. Azevedo (Eds.), *Psychology of stress: new research* (pp. 49–91). Hauppauge, NY: Nova Science Publishers.
- Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology*, 86(3), 389–401. <https://doi.org/10.1037/0022-0663.86.3.389>
- McKendrick, R., Parasuraman, R., Murtza, R., Formwalt, A., Baccus, W., Paczynski, M., & Ayaz, H. (2016). Into the wild: Neuroergonomic differentiation of hand-held and augmented reality wearable displays during outdoor navigation with functional near infrared spectroscopy. *Frontiers in Human Neuroscience*, 10. <https://doi.org/10.3389/fnhum.2016.00216>
- Moyes, J., Sari-sarraf, N., & Gilbert, S. J. (2019). Characterising monitoring processes in event-based prospective memory: evidence from pupillometry. *Cognition*, 184, 83–95. <https://doi.org/10.1016/j.cognition.2018.12.007>
- Myrtek, M., Deutschmann-Janicke, E., Strohmaier, H., Zimmermann, W., Lawrenz, S., Brügger, G., & Müller, W. (1994). Physical, mental, emotional, and subjective workload components in train drivers. *Ergonomics*, 37(7), 1195–1203. <https://doi.org/10.1080/00140139408964897>
- Nathoo, F. S., & Masson, M. E. J. (2016). Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology*, 72, 144–157. <https://doi.org/10.1016/j.jmp.2015.03.003>
- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York, NY, US: W H Freeman/Times Books/ Henry Holt & Co.
- Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (1st ed., pp. 7–15). New York: Lawrence Erlbaum Associates Inc.
- Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7, 44–64.
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A

- cognitive-load approach. *Journal of Educational Psychology*, 84(4), 429–434.
<https://doi.org/10.1037/0022-0663.84.4.429>
- Paas, F. G. W. C., Tuovinen, J., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(2), 63–71.
- Paas, Fred G. W. C., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, 86(1), 122–133. <https://doi.org/10.1037/0022-0663.86.1.122>
- Paas, Fred G. W. C., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79(1), 419–430.
<https://doi.org/10.2466/pms.1994.79.1.419>
- Palinko, O., & Kun, A. (2012). Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 413–416. <https://doi.org/10.1145/2168556.2168650>
- Palinko, Oskar, & Kun, A. L. (2012). Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, 413. <https://doi.org/10.1145/2168556.2168650>
- Palinko, Oskar, Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*, 141. <https://doi.org/10.1145/1743666.1743701>
- Parmentier, D. D., Van, B. B., Detand, J., & Saldien, J. (2019). Design for assembly meaning: a framework for designers to design products that support operator cognition during the assembly process. *Cognition, Technology & Work*. <https://doi.org/10.1007/s10111-019-00588-x>
- Peavler, W. S. (1974). Pupil size, Information overload and performance difference. *Psychophysiology*, 11, 559–566.
- Peysakhovich, V., Causse, M., Scannella, S., & Dehais, F. (2015). Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort. *International Journal of Psychophysiology*, 97(1), 30–37. <https://doi.org/10.1016/j.ijpsycho.2015.04.019>
- Pillay, H. K. (1997). Cognitive load and assembly tasks: Effect of instructional formats on learning assembly procedures. *Educational Psychology*, 17(3), 285–299.
<https://doi.org/10.1080/0144341970170304>
- Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults, 47, 560–569. <https://doi.org/10.1111/j.1469-8986.2009.00947.x>
- Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology*, 60(2), 211–229.
<https://doi.org/10.1080/17470210600673818>
- Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied*, 9(2), 119–137. <https://doi.org/10.1037/1076-898X.9.2.119>
- Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 185–218). Amsterdam: North-Holland: Elsevier Science Publishers B.V.
- Richardson, M., Jones, G., & Torrance, M. (2004). Identifying the task variables that influence perceived object assembly complexity. *Ergonomics*, 47(9), 945–964.
<https://doi.org/10.1080/00140130410001686339>
- Richardson, M., Jones, G., Torrance, M., & Baguley, T. (2006). Identifying the task variables that predict object assembly difficulty. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(3), 511–525. <https://doi.org/10.1518/001872006778606868>
- Richstone, L., Schwartz, M. J., Seideman, C., Cadeddu, J., Marshall, S., & Kavoussi, L. R. (2010). Eye metrics as an objective assessment of surgical skill. *Annals of Surgery*, 252(1), 177–182.
<https://doi.org/10.1097/SLA.0b013e3181e464fb>
- Rosenthal, R. (1994). Parametric measures of effect size. In C. H. & L. V. Hedges (Eds.), *The*

- handbook of research synthesis (pp. 231–244). New York, NY: Russell Sage Foundation.
- Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and Workload Profile methods. *Applied Psychology*, 53(1), 61–86. <https://doi.org/10.1111/j.1464-0597.2004.00161.x>
- Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, 35(11), 991–1009. <https://doi.org/10.1016/j.ergon.2005.04.005>
- Schneider, B. A., Avivi-Reich, M., & Mozuraitis, M. (2015). A cautionary note on the use of the Analysis of Covariance (ANCOVA) in classification designs with and without within-subject factors. *Frontiers in Psychology*, 6(APR), 1–12. <https://doi.org/10.3389/fpsyg.2015.00474>
- Serino, S., Matic, A., Giakoumis, D., Lopez, G., & Cipresso, P. (2016). Pervasive computing paradigms for mental health: 5th International Conference, MindCare 2015, Milan, Italy, september 24–25, 2015 Revised selected papers. *Communications in Computer and Information Science*, 604, 13–22. <https://doi.org/10.1007/978-3-319-32270-4>
- Shalin, V. L., Prabhu, G. V., & Helander, M. G. (1996). A cognitive perspective on manual assembly. *Ergonomics*, 39(1), 108–127. <https://doi.org/10.1080/00140139608964438>
- Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(6), 679–692. <https://doi.org/10.1002/wcs.1323>
- Smallwood, J., Brown, K. S., Baird, B., Mrazek, M. D., Franklin, M. S., & Schooler, J. W. (2012). Insulation for daydreams: A role for tonic norepinephrine in the facilitation of internally guided thought. *PLoS ONE*, 7(4), e33706. <https://doi.org/10.1371/journal.pone.0033706>
- Stinissen, J. (1977). Revised minnesota paper form board test. Vorm AB - Leuvense aanpassing van vorm MA en MB, Likert, R. & Quasha, W. (pp. 1–7). Amsterdam: Swets & Zeitlinger, B.V.
- Stork, S., & Schubö, A. (2010). Human cognition in manual assembly: Theories and applications. *Advanced Engineering Informatics*, 24(3), 320–328. <https://doi.org/10.1016/j.aei.2010.05.010>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(1), 257–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- Sweller, John. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Sweller, John. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load, 22, 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Sweller, John, van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 25–296. <https://doi.org/http://dx.doi.org/10.1023/A:1022193728205>
- Thorvald, P., Lindblom, J., & Andreasson, R. (2017). CLAM - A method for cognitive load assessment in manufacturing. *Advances in Transdisciplinary Engineering*, 6(March 2018), 114–119. <https://doi.org/10.3233/978-1-61499-792-4-114>
- Truschzinski, M., Betella, A., Brunnett, G., & Verschure, P. F. M. J. (2018). Emotional and cognitive influences in air traffic controller tasks: An investigation using a virtual environment, 69, 1–9. <https://doi.org/10.1016/j.apergo.2017.12.019>
- Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., & Jung, T.-P. (2007). Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviation, Space, and Environmental Medicine*, 78, B176–B185.
- Van Acker, B. B., Conradie, P., Vlerick, P., & Saldien, J. (2019). Employee acceptability of wearable mental workload monitoring in industry 4.0: A pilot study on motivational and contextual framing. In *Proceedings of the 22nd International Conference on Engineering Design (ICED19)*. Delft, The Netherlands. <https://doi.org/https://doi.org/10.1017/dsi.2019.216>
- Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology & Work*, 20(3), 351–365. <https://doi.org/10.1007/s10111-018-0481-3>
- van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin and Review*, 1–11. <https://doi.org/10.3758/s13423->

- Van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, 41(2), 167–174. <https://doi.org/10.1111/j.1469-8986.2003.00148.x>
- van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil dilation co-varies with memory strength of individual traces in a delayed response paired-associate task. *PLoS ONE*, 7(12), e51134. <https://doi.org/10.1371/journal.pone.0051134>
- Vansteenkiste, P., Zeuwts, L., Maarseveen, M. Van, Cardon, G., Savelsbergh, G., & Lenoir, M. (2017). The implications of low quality bicycle paths on the gaze behaviour of young learner cyclists. *Transportation Research Part F: Traffic Psychology and Behaviour*, 48, 52–60. <https://doi.org/10.1016/j.trf.2017.04.013>
- Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology*, 52(1), 23–36. <https://doi.org/10.1016/j.ijpsycho.2003.12.003>
- Vidulich, M. A., & Tsang, P. S. (1986). Techniques of subjective workload assessment: A comparison of SWAT and the NASA-Bipolar methods. *Ergonomics*, 29(11), 1385–1398. <https://doi.org/10.1080/00140138608967253>
- Wang, W., Li, Z., Wang, Y., & Chen, F. (2013). Indexing cognitive workload based on pupillary response under luminance and emotional changes. *Proceedings of the 2013 International Conference on Intelligent User Interfaces - IUI '13*, 247. <https://doi.org/10.1145/2449396.2449428>
- Wanyan, X., Zhuang, D., Lin, Y., Xiao, X., & Song, J. W. (2018). Influence of mental workload on detecting information varieties revealed by mismatch negativity during flight simulation. *International Journal of Industrial Ergonomics*, 64, 1–7. <https://doi.org/10.1016/j.ergon.2017.08.004>
- Wickens, C. D. (2017). Mental workload: assessment, prediction and consequences. In L. Longo & M. C. Leva (Eds.), *Human mental workload: models and applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, revised selected papers* (pp. 18–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-61061-0_2
- Wilson, G. F., & Russell, C. A. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(3), 381–389. <https://doi.org/10.1518/hfes.45.3.381.27252>
- Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2014). State of science: Mental workload in ergonomics. *Ergonomics*, 58(1), 1–17. <https://doi.org/10.1080/00140139.2014.956151>